

Section 3 Inference in Simple Regression

Having derived the probability distribution of the OLS coefficients under assumptions SR1–SR5, we are now in a position to make inferential statements about the population parameters: hypothesis tests and confidence intervals.

(Confidence) Interval estimators

- What a confidence interval means: The confidence limits are random variables, the parameter is not. Under the assumptions we have made about the model, 95% of the time our (random) confidence interval will include the actual parameter value.
 - Confidence interval is a pair of random variables β_l and β_u such that $\Pr[\beta_l \leq \beta \leq \beta_u] = 0.95$ or another specified confidence level.

- Under our assumptions, $b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}\right)$

- This is true always if SR6 (normality) is satisfied
- It is true asymptotically if SR6 is not valid but the actual distribution of e has finite fourth moments
- If we know σ^2 then we can convert b_2 to a “standard normal” variable by subtracting its expected value and dividing by its standard deviation:

- $$Z = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2}} \sim N(0,1)$$

- From properties of standard normal, we know that $\Pr[-1.96 \leq Z \leq 1.96] = 0.95$.
- With a little algebra:

$$\Pr\left[-1.96 \leq \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2}} \leq 1.96\right] = 0.95$$

$$\Pr\left[-1.96 \times \sqrt{\sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2} \leq b_2 - \beta_2 \leq 1.96 \times \sqrt{\sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2}\right] = 0.95$$

$$\Pr\left[b_2 - 1.96 \times \sqrt{\sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2} \leq \beta_2 \leq b_2 + 1.96 \times \sqrt{\sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2}\right] = 0.95.$$

- This is the 95% interval estimate (usually called a confidence interval) for β_2 .

- We can't use the formula above unless we know σ^2 , which we normally don't.
 - If we replace σ^2 by s^2 , then b_2 follows a t distribution with $N - 2$ degrees of freedom rather than the normal distribution.
 - $$t = \frac{b_2 - \beta_2}{\sqrt{s^2 / \sum_{i=1}^N (x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{s.e.(b_2)} \sim t_{(N-2)}$$
 - If t_c is the 5% two-tailed critical value for this t distribution, then $\Pr[b_2 - t_c \times s.e.(b_2) \leq \beta_2 \leq b_2 + t_c \times s.e.(b_2)] = 0.95$
 - Explain how to find the critical value, both theoretically and in the tables.
 - Stata (and some other packages) prints out these confidence limits based on assumptions SR1–SR6.
- Interval estimators in Monte Carlo simulations
 - What if we don't know the distribution of our estimator from theory?
 - Can't read critical values from the t or N tables
 - Can get them from simulations
 - Monte Carlo simulation gives us, say, 10,000 replications of the sampling experiment from the model that we assume generated our data
 - We have used this to assess the mean and standard deviation (to estimate the standard error) of our estimator
 - We can also look at the quantiles of this estimated distribution to get estimated critical values:
 - In 10,000 replications of the test statistic, 250 will lie in the upper 2.5% tail and 250 in the lower 2.5% tail.
 - The 97.5% and 2.5% quantiles (250th and 9,750th values in rank order) give estimates of the appropriate critical values
 - In Stata, the centile `b`, `centile(2.5 97.5)` calculates these values for an estimator whose M simulated values are stored in `b`

Hypothesis tests about single coefficients

- The most common test in econometrics is the “ t -test” of the hypothesis that a single coefficient equals zero.
 - This test is printed out for each regression coefficient in Stata and other statistical packages.
 - Depending on the assumptions of the model (and whether they are valid), the “ t -statistic” may or may not follow Student's t distribution.
- General form for calculating a t -statistic is $t = \frac{b_k - c}{s.e.(b_k)}$, where c is the hypothetical value (often zero) that we are testing against and $s.e.$ is the standard error of the coefficient estimator b_k .

- This test statistic is useful because we know its distribution under the null hypothesis that $\beta_k = c$.
 - Thus we can determine how likely or unlikely it is that we would observe the current sample if the null hypothesis is true.
 - This allows us to control the Type I error at significance level α .
- Using the t -statistic to test $H_0: \beta_k = c$ against the two-sided alternative $H_1: \beta_k \neq c$
 - Note that hypothesis to be tested is always expressed in terms of the *actual* coefficient, not the estimated one.
 - Use the formula above to calculate the t statistic. (Stata will print out b and its standard error, and also the t value corresponding to $c = 0$.)
 - If the absolute value of the calculated t value is greater than the critical value, then reject the null.
 - Do example including looking up critical value.
 - Alternatively, we can compute the probability (p) value associated with the test: the probability that an outcome at least this inconsistent with the null hypothesis would occur if the null is indeed true.

$$\begin{aligned}
 p &= \Pr_{H_0} \left[|b_2 - c| > |b_2^{act} - c| \right] \\
 &= \Pr_{H_0} \left[\frac{|b_2 - c|}{s.e.(b_2)} > \frac{|b_2^{act} - c|}{s.e.(b_2)} \right], \text{ where the "act" refers to the actual} \\
 &= \Pr_{H_0} (|t| > |t^{act}|).
 \end{aligned}$$

observed/calculated value

- If we know the distribution of the t statistic, then we can calculate the last probability from tables.
 - Under assumptions SR1–SR5, the t statistic will be asymptotically normal.
 - With SR6 it is t with $N - 2$ degrees of freedom in small samples.
 - Stata calculates the p value associated with the null hypothesis $\beta = 0$ using the t distribution
 - Show diagram corresponding to HGL’s Figure 3.2 on page 103: For given $|t|$, show how to calculate p value.
 - On same diagram show *critical values* for test at given level of significance, and how to decide the result of the test
 - Note 1.96 as two-tailed 5% critical value for normal distribution.
 - Then show the symmetry: the p value is the smallest significance level at which the null hypothesis can be rejected.
- Critical values in Monte Carlo simulations
 - As with interval estimators, we can use the estimated quantiles of a statistic from a Monte Carlo simulation to calculate critical values

- This is particularly useful when we don't know the true distribution of the estimator under the null hypothesis
- The Monte Carlo simulation is *always conducted under the assumption that the null hypothesis is true*
- One-tailed test such as $H_0: \beta_2 = c$, $H_1: \beta_2 < c$. (Or $H_0: \beta_2 \geq c$)
 - Same basic procedure, but in this case we concentrate the entire rejection region in one tail of the distribution.
 - We reject the null if and only if $\Pr[t < t^{act}] < \text{critical value}$ (ignoring right tail of distribution) and fail to reject for any positive t value no matter how large.
 - Other direction if H_1 is $\beta_1 > c$: Fail to reject null for any negative value of t and reject when $\Pr[t < t^{act}] > \text{critical value}$.
- Present some examples of regressions and practice with tests of $\beta = 0$ and $\beta = \text{other values}$.
 - Good (multiple regression) example with lots of different significance levels:
 - reg gpoints irldr satv100 satm100 taking if freshman
 - Can do just taking to get almost significant example for simple regression

Testing linear combinations of parameters

- What if we want an interval estimator or hypothesis test for the value of y when $x = 100$?
- This would be an estimator for $\beta_1 + 100\beta_2$.
- The natural estimator is $b_1 + 100b_2$
- What is the distribution of $b_1 + 100b_2$?
 - $b_1 + 100b_2$ is a linear function of b_1 and b_2 , so it is normal (or t) if they are
 - $E(b_1 + 100b_2) = E(b_1) + 100E(b_2) = \beta_1 + 100\beta_2$ so it is unbiased (under assumptions)
 - $\text{var}(b_1 + 100b_2) = \text{var}(b_1) + 10,000 \text{var}(b_2) + 200 \text{cov}(b_1, b_2)$
 - We can approximate these variances and covariance by their sample estimators, and use the result to calculate a t statistic.
- Can also do hypothesis test of such linear combinations of coefficients.
- Stata command: lincom

Prediction in the simple regression model

- One of the most common tasks for which we use econometrics is conditional prediction or forecasting.
 - We want to answer the question “What would y be if x were some value x_0 ?”
- This is exactly the same problem we discussed above in estimating the distribution of $b_1 + 100b_2$, which is the OLS prediction of y for $x = 100$.

- OLS prediction: $\hat{y}_0 = b_1 + b_2 x_0$
 - Because $E(e) = 0$, we “predict” it to be zero.
 - Might not do that if we had information about the error term corresponding to our prediction.
 - Note that we are assuming x_0 to be given.
 - Secondary prediction problem occurs if we must also predict x .
- Forecast error (prediction error) is $f \equiv y_0 - \hat{y}_0$.
 - $f = \beta_1 + \beta_2 x_0 + e_0 - (b_1 + b_2 x_0)$
 - $= (\beta_1 - b_1) + (\beta_2 - b_2) x_0 + e_0$
 - $E(f) = 0$ because b is unbiased, so OLS predictor is unbiased
 - OLS predictor is BLUP based on BLUE OLS estimator
- What is the variance of \hat{y}_0 or, equivalently, the variance of f ?
 - $f = (\beta_1 - b_1) + (\beta_2 - b_2) x_0 + e_0$
 - $\text{var}(f) = \text{var}(\hat{y}_0) = E(y_0 - \hat{y}_0)^2$
 - $= \text{var}(b_1) + x_0^2 \text{var}(b_2) + 2x_0 \text{cov}(b_1, b_2) + \text{var}(e_0)$.
 - For simple regression under homoskedasticity,

$$\begin{aligned}
 \text{cov}(\mathbf{b}) &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix}^{-1} \\
 &= \frac{\sigma^2}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix} \\
 &= \frac{\sigma^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} \begin{bmatrix} \sum_{i=1}^N x_i^2 & -\sum_{i=1}^N x_i \\ -\sum_{i=1}^N x_i & N \end{bmatrix}.
 \end{aligned}$$

o So

$$\begin{aligned} \text{var}(\hat{y}_0) &= \sigma^2 \left[1 + \frac{Nx_0^2 - 2x_0N\bar{x} + \sum_{i=1}^N x_i^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} \right] \\ &= \sigma^2 \left[1 + \frac{\sum_{i=1}^N x_i^2 - N\bar{x}^2 + Nx_0^2 - 2x_0N\bar{x} + N\bar{x}^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} \right] \\ &= \sigma^2 \left[1 + \frac{\sum_{i=1}^N (x_i - \bar{x})^2 + N(x_0 - \bar{x})^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} \right] \\ &= \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \right]. \end{aligned}$$

o Prediction error is smaller for:

- Smaller error variance
- Larger sample size (through both second and third terms)
- Greater sample variation in x
- Observations closer (X) to the mean

▪ With SR6 (normality) or asymptotically under more general assumptions,

$$\frac{f}{\sqrt{\text{var}(f)}} \sim N(0,1), \text{ because } f \text{ is a linear function of normal variables with mean 0.}$$

▪ We usually don't know σ^2 , so we must replace it with s^2 . This makes the distribution t_{N-2} rather than normal.

▪ Interval estimate for \hat{y}_0 is $\Pr[\hat{y}_0 - t_c \times s.e.(\hat{y}_0) \leq y_0 \leq \hat{y}_0 + t_c \times s.e.(\hat{y}_0)] = 1 - \alpha$, where t_c is the $\alpha/2$ critical value of the t_{N-2} distribution.

▪ **Stata predict command**

o You get a new variable with the predictions for whatever sample you set in the if clause. (Will do whole sample by default, even if the regression was run on a subsample.)

o What you get is driven by the options:

- Default is **xb**, which gives fitted values
- **resid** gives residuals
- **stdp** gives $\sqrt{\widehat{\text{var}}(\hat{y}_i)}$

$$\text{▪ } \mathbf{stp} \text{ gives } \sqrt{\widehat{\text{var}}(y_i - \hat{y}_i)} = \sqrt{\widehat{\text{var}}(y_i) + \widehat{\text{var}}(\hat{y}_i)} = \sqrt{\widehat{\text{var}}(\hat{y}_i) + \hat{\sigma}^2}$$

Measuring goodness of fit

- It is always of interest to measure how well our regression line fits the data. There are several measures that are commonly reported.

- Sum of Squares due to Errors = $SSE = \sum_{i=1}^N \hat{\epsilon}_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$, with $\hat{y}_i \equiv y_i - b_1 - b_2 x_i$.

- $SST = \text{Sum of Squares Total} = \sum_{i=1}^N (y_i - \bar{y})^2$.

- $SSR = \text{Sum of Squares due to Regression} = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2$

- Warning about notation: some books use SSR for “sum of squared residuals” and SSE to mean “sum of squares explained.”

- Fundamental regression identity: $SST = SSR + SSE$. Works due to the enforced independence of \hat{y} and $\hat{\epsilon}$. See Appendix 4B.

- **Standard error of the estimate (regression):** This is our estimate of the standard deviation of the error term.

- $SEE = s_{\hat{\epsilon}} = \sqrt{s_{\hat{\epsilon}}^2} = \sqrt{\frac{1}{N-2} \sum_{i=1}^n \hat{\epsilon}_i^2} = \sqrt{\frac{SSE}{N-2}}$.

- Standard error of regression is often (as in Stata) called *root mean squared error* or RMSE.

- **Coefficient of determination: R^2**

- The R^2 coefficient measures the fraction of the variance in the dependent variable that is explained by the covariation with the regressor. It has a range of $[0, 1]$, with $R^2 = 0$ meaning no relationship and $R^2 = 1$ meaning a perfect linear fit.

- $R^2 \equiv \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{N-2}{N-1} \frac{s_{\hat{\epsilon}}^2}{s_y^2}$.

- R^2 is approximately the square of the sample correlation coefficient between y and \hat{y} .

Specification issues

- **Scaling**

- Does it matter how we scale the x and y variables?
- If we **add or subtract** a constant from either x or y , all that is affected is the intercept term b_1 . Since we are not usually very interested in the value of the intercept, this is usually meaningless.
- If we **multiply x** by a constant, the slope estimate b_2 will be divided by the same constant, as will its standard error, leaving the t statistic unchanged. The estimated intercept is unchanged, as are the residuals, SEE, and R^2 .

- If we **multiply** y by a constant, the slope and intercept estimates will both be multiplied by the same constant, as will their standard errors (leaving the t statistics unchanged) and the SEE.
- None of these transformations has any effect on R^2 .
- **Nonlinear models**
 - We can easily replace either x or y with powers or logs of the original variables without complicating the estimation.
 - What changes is the shape of the relationship between the levels of x and y and the interpretation of the coefficients.
 - HGL Fig 4.5 on p. 142 and Table 4.1 on p. 143

Table 4.1 Some Useful Functions, their Derivatives, Elasticities and Other Interpretation

Name	Function	Slope = dy/dx	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x) \frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2) \frac{x}{y}$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-Linear	$\ln(y) = \beta_1 + \beta_2 x$ or, a 1 unit change in x leads to (approximately) a 100 $\beta_2\%$ change in y	$\beta_2 y$	$\beta_2 x$
Linear-Log	$y = \beta_1 + \beta_2 \ln(x)$ or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$

142 PREDICTION, GOODNESS-OF-FIT, AND MODELING ISSUES

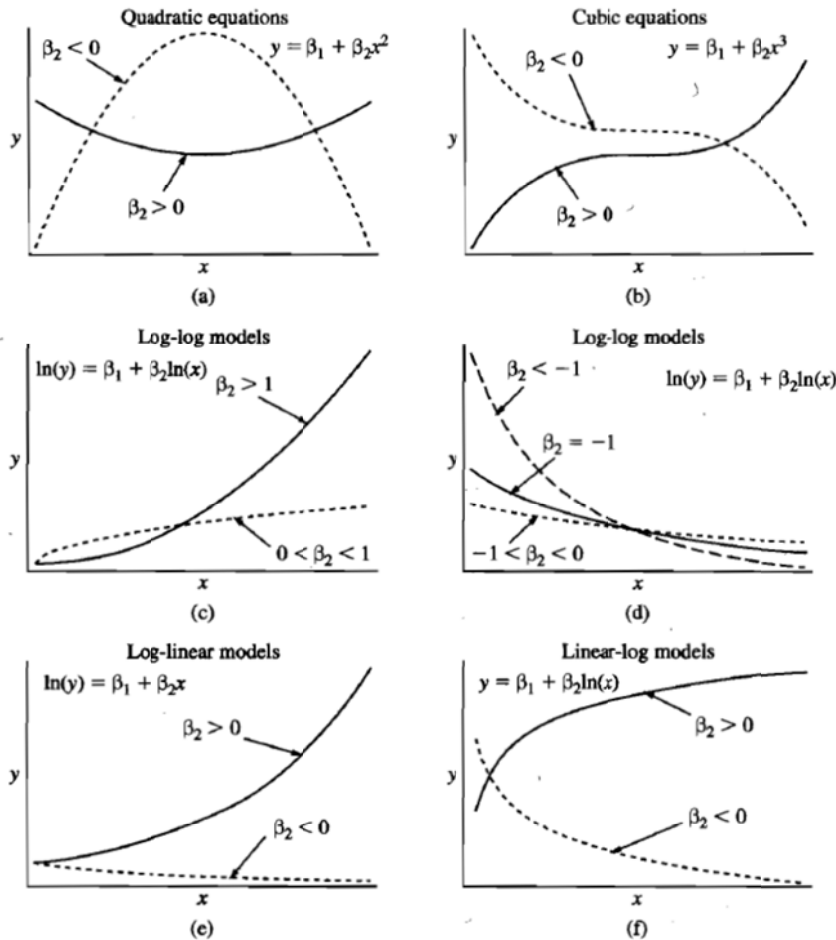
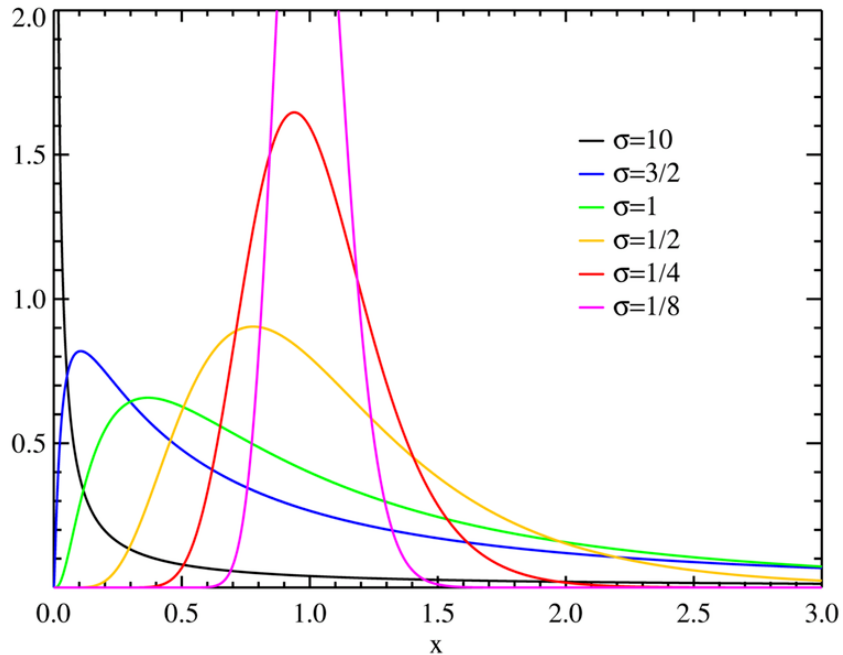


FIGURE 4.5 Alternative functional forms.

Log-based models

- Many econometric models are specified in log term.
 - Most economic variables are non-negative, so we don't need to worry about negative values. (Though many can be zero.)
 - $d(\ln x) = dx/x$ = the percentage change in x , so the interpretation of coefficients and effects is useful and easy.
 - The log-log model is a constant-elasticity specification with the coefficient being read directly as an elasticity.
 - Shape of log functions is often reasonable:
 - Shies away from axes
 - Monotonic with diminishing returns
- **Log of regressor only** (“linear-log” model)
 - $y_i = \beta_1 + \beta_2 \ln x_i + e_i$.
 - Change of 1% in x changes $\ln x$ by about 0.01 and thus leads to about a $0.01\beta_2$ unit absolute change in y .
 - If x increases by $z\%$, this means it is $1 + z/100$ times as large, which means that its log is $\ln x_0 + \ln(1 + z/100)$. If z is small (say, less than 20%) then the approximation is reasonable close. However, you may want to do exact calculations for formal work.
 - Partial effect in levels is $\frac{\partial y}{\partial x} = \beta_2 \frac{1}{x}$, which is monotonically increasing or decreasing (depending on sign of β_1) but slope goes to zero as x gets large.
- **Log of dependent variable only** (“log-linear” model)
 - $\ln y_i = \beta_1 + \beta_2 x_i + e_i$
 - Note that $y_i = e^{\beta_1 + \beta_2 x_i + e_i}$, so this is clearly a different error term than when y is not in log terms.
 - Change of z units in x changes $\ln y$ by $\beta_2 z$ units, so it changes y by about $100\beta_2 z$ percent.
 - The same approximation issues applies here. The increase of $\beta_1 x$ units in $\ln y$ means that y increases by a factor of $e^{\beta_2 z}$, which is approximately $1 + \beta_2 z$ for small values of z . For larger values of z and for more formal work, it is best to calculate the exponential directly.
 - Partial effect in levels is $\frac{\partial y}{\partial x} = \frac{\partial \ln y}{\partial x} \frac{\partial y}{\partial \ln y} = \beta_2 / \frac{\partial \ln y}{\partial y} = \beta_2 y$.
 - Alternatively, $\frac{\partial y}{\partial x} = \beta_2 e^{\beta_1 + \beta_2 x + e} = \beta_2 y$.
 - Partial effect is increasing in absolute value as y increases. (Note that y must always be positive in this model.)

- **Log of both regressor and dependent variable** (“log-log model”)
 - $\ln y_i = \beta_1 + \beta_2 \ln x_i + e_i$.
 - Also implies that $y_i = e^{\beta_1 + \beta_2 \ln x_i + e_i} = e^{\beta_1} x_i^{\beta_2} e^{e_i} = \alpha_0 x_i^{\beta_2} \nu_i$, where $\alpha_0 \equiv e^{\beta_1}$, $\nu_i \equiv e^{e_i}$.
 - The Cobb-Douglas function takes this form (with a multiplicative error ν , usually assumed to be log-normally distributed).
 - Change of 1% in x changes $\ln x$ by about 0.01, which changes $\ln y$ by about $0.01\beta_2$, which changes y by about $\beta_2\%$. (Both of the approximation caveats above apply here.)
 - Thus, **β_2 is the point elasticity of y with respect to x .**
 - This makes log-log a popular function form.
 - Partial effect in levels is $\frac{\partial y}{\partial x} = \frac{\partial y}{\partial \ln y} \frac{\partial \ln y}{\partial \ln x} \frac{\partial \ln x}{\partial x} = \beta_2 \frac{y}{x}$.
 - Alternatively, $\frac{\partial y}{\partial x} = e^{\beta_1 + \beta_2 \ln x + e} \beta_2 \frac{\partial \ln x}{\partial x} = \beta_2 \frac{y}{x}$.
 - Partial effect is constant in elasticity terms, but varies with y and x in level terms.
- Which log model to choose? Theory vs. let data decide?
 - Theory may suggest that percentage changes are more important than absolute changes for one or both variables.
 - Income is often logged if we think that a doubling of income from \$50,000 to \$100,000 would be associated with the same change in other variables as a doubling from \$100,000 to \$200,000 (rather than half as much).
 - As suggested by the previous example, logging a variable scales down extreme values. If most of the sample variation is between \$20,000 and \$100,000 (with mean \$50,000 and standard deviation \$30,000), but you have a few values of \$500,000 for income, these are going to be 15 standard deviations above the mean in level terms but much less in log terms.
 - The log of 500,000 is only $\ln(10)=2.3$ units larger than the log of 50,000. The standard deviation of the log would probably be in the range of 0.6 or so, so the highly deviant observations would be less than 4 standard deviations above the mean instead of 15.
 - Since we often want our variables to be normally distributed, we might try to decide whether the variable is more likely to be normally or log-normally distributed.



- Log-normal distribution
- Note that the various log models are **not nested** with one another or with the linear or polynomial models, so t tests cannot discriminate between them.
- We can use R^2 to compare models if only if the dependent variable is the same:
 - Linear model with linear-log model
 - Log-linear model with log-log model
- **Box-Cox model** nests log and linear terms for both dependent and independent variables in a nonlinear model.
 - Can estimate Box-Cox model and test hypothesis that the relationship is linear or log.
 - Box-Cox transformation is $B(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln x, & \text{if } \lambda = 0. \end{cases}$
 - This is a continuous function that equals $x - 1$ if $\lambda = 1$ and $\ln x$ if $\lambda = 0$.
 - We can do a nonlinear regression of $B(y, \lambda_y)$ on $B(x, \lambda_x)$ and test the two λ values to see whether they are zero or one to determine whether a linear or log specification is preferred for both variables.
- **Prediction in log and other y -transformed models**
 - If we run the regression $g(y) \equiv z = \beta_1 + \beta_2 x + e$, how would we predict y ?
 - This problem usually arises when $g(y) = \ln y$.

- We can predict $\ln y$ by $\widehat{\ln y} = b_1 + b_2 x + E(e) = b_1 + b_2 x$. But $E(y) = E(e^{\ln y}) > e^{E(\ln y)} = e^{\widehat{\ln y}}$. So the true values will be larger than our naïve predictions: We will underpredict y consistently.
- The problem is that even if $E(e_i) = 0$, $E(\exp(e_i)) \neq 1$.
 - Consider normally distributed e with mean zero and variance one.
 - The mean of zero translates into $\exp(0) = 1$. So far so good.
 - Now consider two equally likely draws from this distribution at $+1$ and at -1 .
 - The $+1$ draw translates into $\exp(1) = 2.718$.
 - The -1 draw translates into $\exp(-1) = 1/2.718 = 0.368$
 - These two draws average to $(2.718 + 0.368) / 2 = 1.543 > 1$
- If e is normally distributed with variance σ^2 , then $E(e^{e_i}) = e^{\frac{\sigma^2}{2}}$.
 - In that case, we can predict y by $\hat{y}_e = e^{\frac{\sigma^2}{2}} e^{\widehat{\ln y}}$. This is a consistent prediction if the error term is normal.

○ Very effective example using HGL's cps4_small data set

```
. g lwage=log(wage)
```

```
. reg lwage educ
```

Source	SS	df	MS	Number of obs =	1000
Model	60.015841	1	60.015841	F(1, 998) =	216.41
Residual	276.76489	998	.27731953	Prob > F =	0.0000
Total	336.780731	999	.337117849	R-squared =	0.1782
				Adj R-squared =	0.1774
				Root MSE =	.52661

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ	.0904082	.0061456	14.71	0.000	.0783484 .1024681
_cons	1.609444	.0864229	18.62	0.000	1.439853 1.779036

```
. predict lwage_hat
(option xb assumed; fitted values)
```

```
. g wage_hat=exp(lwage_hat)
```

```
. summarize lwage lwage_hat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lwage	1000	2.856988	.5806185	.6780335	4.335852
lwage_hat	1000	2.856988	.2451039	1.609444	3.508018

```
. summarize wage wage_hat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1000	20.61566	12.83472	1.97	76.39
wage_hat	1000	17.94095	4.540388	5.000033	33.38202

```
. display exp(.27731953/2)
1.1487332
```

```
. gen wage_hat_c=wage_hat*exp(.27731953/2)
```

```
. summarize wage wage_hat wage_hat_c
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1000	20.61566	12.83472	1.97	76.39
wage_hat	1000	17.94095	4.540388	5.000033	33.38202
wage_hat_c	1000	20.60937	5.215694	5.743704	38.34704

- In the non-normal case, we can use a simple regression to calculate the appropriate adjustment factor:
 - Run a regression of $y_i = \gamma e^{\widehat{\ln y}_i}$, which is a bivariate regression without a constant term.
 - Then adjust the predictions to get $\hat{y}_c = \hat{\gamma} e^{\widehat{\ln y}}$, which, for the sample observations, are just the predicted values from the auxiliary regression.
 - Can't do a convenient interval predictor because \hat{y}_c is not normal or t distributed.

Using residuals

- All regression software will have a way to scatter-plot the actual and fitted values or the residuals against another variable (x is often most useful).
 - Don't put residual plot and actual/fitted plot on same diagram because of scaling.
- Residuals tell you what you are missing in your regression:
 - None of us is perfect: we learn from our mistakes
 - Our regression is not perfect if $R^2 < 1$
 - We learn from our residuals
 - Functional form: if there is obvious curvature in the actual vs. fitted values, you may need a nonlinear form
 - Heteroskedasticity: if the variance of the residuals seems to be related to x or another variable, then you may need to correct for it.
 - How would you tell this from residual plot?
 - Outliers: Are there specific observations that are far from the normal pattern?

- If so, they may indicate that one or more observations do not follow the same model (Assumption 0).
 - Or they may suggest an additional explanatory variable that affected y in those observations.
 - Or they might be coding errors in the data set.
- Are residuals normally distributed?
 - If error term is normal, then residuals should be.
 - Jarque-Bera test
 - $$JB = \frac{N}{6} \left(S^2 + \frac{(K-3)^2}{4} \right)$$
 - Tests whether skewness and kurtosis of variable match the zero, three expected in normal distribution.
 - There are other tests as well.